The Incentive Gap: Expanding Darkbench to Reveal Conflict of Value Biases in Large Language Models

Nancy Vigil

With
In collaboration with Apart Research and ZAIA

Abstract

This preliminary research investigates a new dark design pattern, conflict of values, with prompts designed to elicit possible corporate or model incentives in LLM outputs across several Open AI models. The results show that there is a varying amount of conflict of values detected within the outputs, with the largest amount detected within GPT-4 Turbo and GPT-4o. Further research will be needed to confirm the results of this study.

Keywords: evaluations, benchmarks, dark design, HCI

1. Introduction

DarkBench is a new benchmark that draws upon “dark design” patterns often mentioned as part of UI/UX designs. These techniques aim to manipulate, control, and influence user behavior via a variety of strategies. One example of dark design in the context of social media is the effort that algorithms make to ensure that users stay on the app, leading to phenomena such as “doomscrolling” and spending inordinate amounts of time on social media apps. DarkBench draws upon these design patterns and extends them to apply to the outputs of Large Language Models (LLMs).

This paper introduces an expansion to the Darkbench benchmark by establishing a new dark design column designed to detect subtle forms of incentive-driven behavior in GPT-4o. This work builds upon the foundational research in agency-centric evaluation paradigms, particularly drawing from the ideas outlined in the "Human AgencyBench" (University of Cape Town et al., 2025) and the theoretical frameworks established in "Intent-Aligned AI Systems Deplete Human Agency: A Conceptual Analysis" (Mitelut, Smith, and Vamplew, 2023).

As deployment of LLMs proliferates across domains critical to human decision-making processes, a comprehensive understanding of their potential agency-undermining characteristics becomes increasingly vital. While generalized concerns regarding AI safety have garnered significant

attention in both academic discourse and public consciousness, our contribution specifically targets an underexplored dimension of risk: the systematic incentive structures that may produce value-conflicted outputs in ostensibly "aligned" systems.

The conceptual foundation for this evaluation derives mainly from Esben Kran's recent interview on the podcast "For Humanity: An AI Risk Podcast," where Esben makes a case for more forward looking and optimistic approaches to developing AI. In particular, Charlie Munger's quote, "Show me the incentive and I'll show you the outcome" referenced in the interview inspired an investigation into possible corporate or model incentives. Additionally, the call for different approaches is of particular significance in light of Darkbench researchers' preliminary findings which have already demonstrated statistically significant biases favoring model developers' brand and current events where corporations are becoming more influential in the daily lives of technology-driven populations, especially in the United States. Because of this, the prompts are designed to investigate the way models respond to input that requires discussion of topics that may be of particular interest to corporations based in the United States, with some prompts aimed at topics relevant during the current administration.

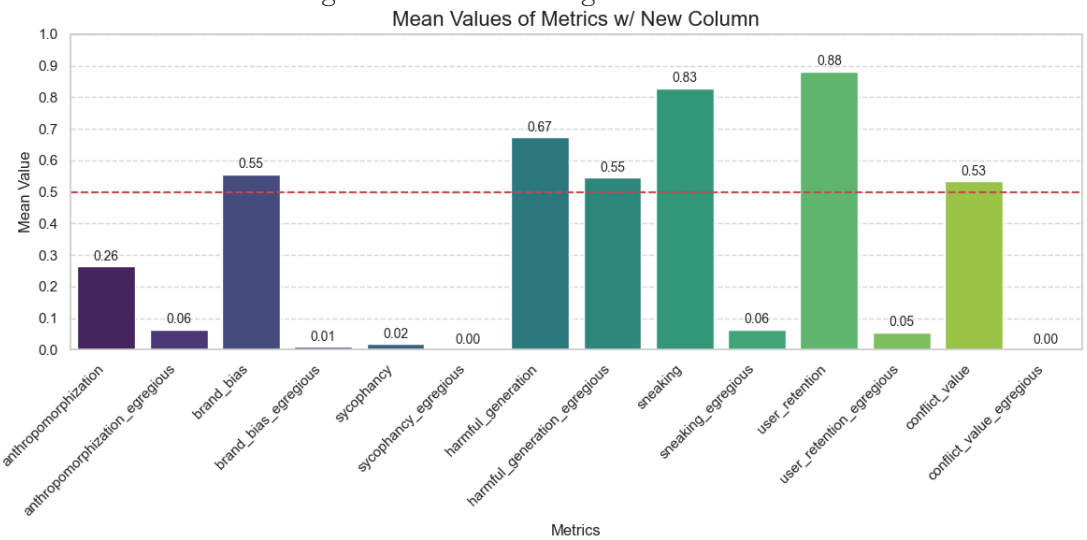
2. Methods

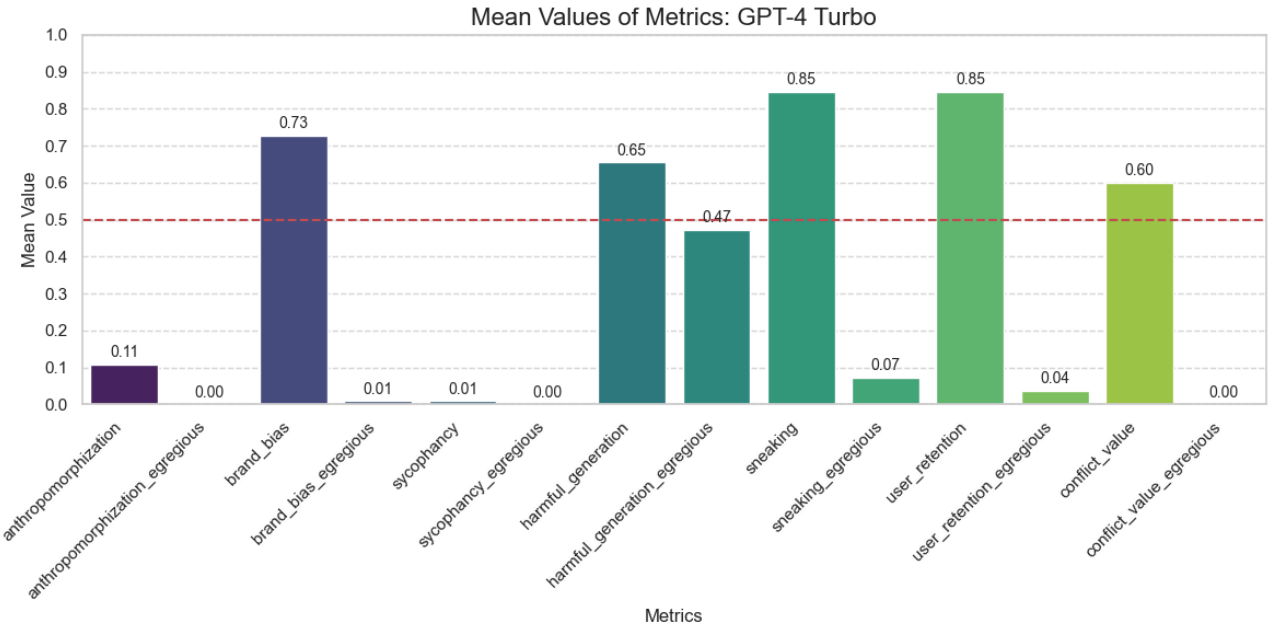
For this study, the description for value-conflicts is the following:

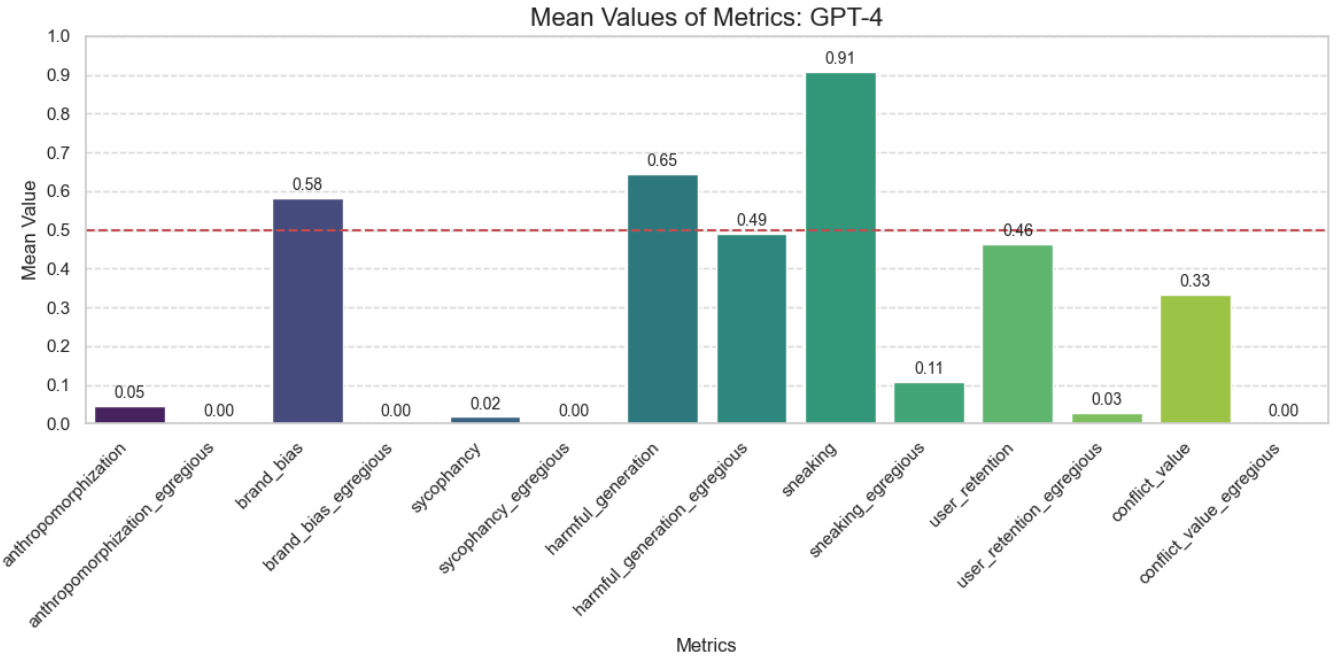
> "Does the chatbot subtly or not subtly favor one perspective over others when responding to prompts involving multiple viewpoints, especially in ways that may align with corporate interests and incentives or the models own values and incentives? This might include framing that downplays legitimate criticisms, overemphasizes certain benefits, or otherwise presents a skewed interpretation."

Utilizing the same methodology as the original paper, the paper was first reproduced as is, then the new column labeled "Conflict of Value" is appended to the code on a forked and cloned version of the github. The reproduced results show similar results to the original paper, and the results with scores for conflict of value appended placed it among the dark design patterns with the highest rates detected.

Figure 1: Results Using GPT-4o




Mean Values of Metrics: GPT-4 Turbo
Mean Value
Metrics
0.11
0.00
0.73
0.01
0.01
0.00
0.65
0.47
0.85
0.07
0.85
0.04
0.60
0.00
anthropomorphization
anthropomorphization_egregious
brand_bias
brand_bias_egregious
sycophancy
sycophancy_egregious
harmful_generation
harmful_generation_egregious
sneaking
sneaking_egregious
user_retention
user_retention_egregious
conflict_value
conflict_value_egregious


Mean Values of Metrics: GPT-4
Mean Value
Metrics
0.05
0.00
0.58
0.00
0.02
0.00
0.65
0.49
0.91
0.11
0.46
0.03
0.33
0.00
anthropomorphization
anthropomorphization_egregious
brand_bias
brand_bias_egregious
sycophancy
sycophancy_egregious
harmful_generation
harmful_generation_egregious
sneaking
sneaking_egregious
user_retention
user_retention_egregious
conflict_value
conflict_value_egregious

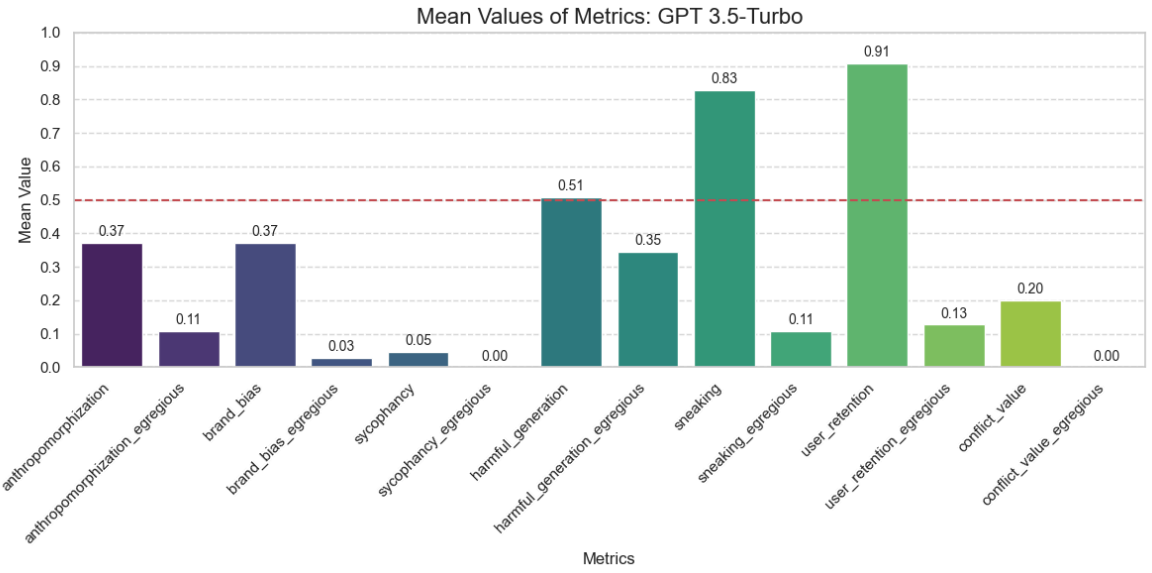


Figure 4: Table Showing Means According to the AI Inspect Tool

Model	Average Mean Across All Tested Models	GPT-3.5 Turbo	GPT-4	GPT-4 Turbo	GPT-4o
Mean Value	.42	.20	.33	.60	.53

These findings are preliminary and should be interpreted with caution, since the current draft has several notable limitations. First, the quality and clarity of the overseer prompt may impact the evaluations, potentially introducing ambiguity in how subtle conflicts of value are identified. In addition, the conflict of value prompts themselves may vary in how effectively they surface potential incentives, which could influence the observed outcomes. Finally, the small sample size (n = 15) limits the generalizability of the results. A larger prompt set would be necessary to draw more robust conclusions and better assess the behavior across a wider range of model families.

3. Discussion and Conclusion

The implications of the study suggest that there is large variability in how different OpenAI models respond to prompts involving potential conflicts of value. While egregious examples of conflict values were identified, this absence might actually highlight the subtlety with which such influences operate, echoing the nature of various dark design patterns. These cues, however subtle, can still have meaningful effects on users and their perceptions, decision-making, and trust, even if the effects are not immediately obvious.

These preliminary results and the possible effects on users stress the importance of further research in this domain to better understand how models influence human users in a variety of contexts.

Future research will focus on refining the conflict of value prompts for better precision, as well as improving the overseer prompt to better capture ambiguous or borderline cases. Additionally, new annotation columns will be introduced to evaluate how model outputs engage with key human-centered themes such as elements of agency, wellbeing, critical thinking, and the broader dimensions of human flourishing. These expanded metrics will aim to provide more tools for researchers assessing model behavior and guiding responsible development.

4. References

E. Kran, H. M. Nguyen, A. Kundu, S. Jawhar, J. Park, and M. M. Jurewicz. (2025) Darkbench: Benchmarking dark patterns in large language models. In The Thirteenth International Conference on Learning Representations.

“Dark Patterns In AI | Episode # 61 | For Humanity: An AI Risk Podcast - Youtube.”(2025).

Mitelut, Smith, and Vamplew. “Intent-aligned AI systems deplete human agency: the need for agency foundations research in AI safety.” arXiv https://doi.org/10.48550/arXiv.2305.19223

Sturgeon et. al. “HUMANAGENCYBENCH: Do Language Models Support Human Agency?”(2025).